A Modern Approach to Predicting CO_2 emissions in Canadian ICE (Internal Combustible Engines)

DATA 621: Fall 2020 Final Project

Group 3: Zach Alexander, Sam Bellows, Donny Lofland, Joshua Registe, Neil Shah, Aaron Zalki

Table of Contents

Abstract 3

Key Words 4

Introduction 5

Literature Review 6

Methodology 7

Experimentation and Results 8

Discussions and Conclusions 12

References 14

Table of Figures

Figure 1 Canadian Energy Flows 5

Figure 2 European ICE emissions (G. Mellios, 2011) 6

Figure 4 - Distribution of Numeric Variables 8

Figure 5 - Quantitative description of Qualitative Variables 9

Figure 6 - Correlation Matrix of all Variables 10

Figure 7 - Numeric Variables vs. CO2 Emissions 10

Figure 8 - Distribution of Transformed Variables 11

Figure 9 - Missing Variables 11

Figure 10 - Test/Train comparison for Model Performance 12

Figure 11 - Variable Importance for Models 13

Abstract

In this paper, we set out to determine what factors play a large role in the amount of CO_2 emissions produced by a vehicle and how we can use these findings to guide policymaking to reduce emissions and combat climate change effectively. The topic is an important one, as climate change can have adverse economic and social effects. More and more governments are attempting to limit their impact on the environment by enacting policies to reduce their carbon footprint. Our main findings are that both *engine size* and *fuel type* are highly correlated with a vehicle's CO_2 emissions. Therefore, limiting engine size and banning certain types of fuels could lead to a decrease in CO_2 emissions. Our secondary finding was that fuel consumption was highly correlated with both emissions and engine size and that enacting road standards for certain fuel consumption levels may help reduce emissions. This is not an unheard-of finding as many governments have already enacted laws or incentives to improve fuel consumption rates in new cars.

Key Words

Carbon-Dioxide, ICE, Emissions, Vehicle, Canada

Introduction

Climate change—the impact of human-made activities on greenhouse gases and their role in changing the climate—has emerged as a top priority for most Canadians in 2019 (Shah, 2019). This is particularly important given that Canada's primary exports are still crude-oil (primarily oil sands), refined products (Irving Oil is one of the largest refineries), and other fossil fuel-related activities, accounting for nearly 10% of the national GDP (Government du Canada, n.d.).

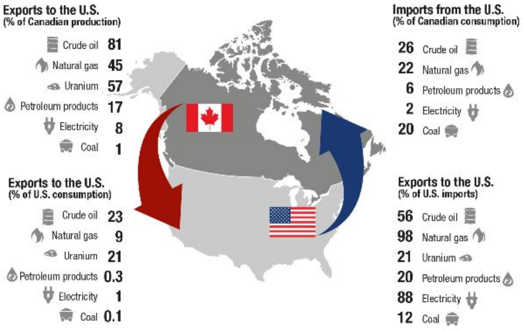


Figure 1 Canadian Energy Flows

While the global scientific community has had consensus since the 1990s, the economic impacts of climate change have been thrust to the forefront, with the ECCC reporting nearly 1.5-23% cost to the Canadian GDP towards the end of the century (Molico, 2019). To mitigate the impact of climate change, Canadian policymakers have adopted the Low Carbon Fuel Standard (LCFS) (Clean Fuel Standard, 2019), mirroring that of California. California has long been on the forefront of clean energy, having maintained its own stricter gasoline standard CARBOB, a zero-emission vehicle (ZEV) standard, and even tighter vehicle emission standards (CAA Section 909), while setting nation-wide vehicle standards and earning the ire of the 2020 EPA administration (Tabuci, 2018). The Low Carbon Fuel Standard is a framework that scores carbon intensity from all energy sources—power, shipping, and vehicles—and seeks to promote low carbon fuel sources.

Given the recentness—at the time of this report's writing, the Canadian LCFS was announced in 2019—there is increasing importance on assessing how much CO_2 the Canadian car fleet emits and what factors are of primary concern. Our team seeks to quantify said carbon dioxide emissions using generalized linear models on a provided vehicle data set. Such a model would help assess Canada's progress towards a lower carbon future and serve as a tool for policymakers in other regions to evaluate similar programs' efficacy.

Literature Review

Emissions are an active area of research, and both the industry and scientific community have a rich history of empirical, policy, and simulations to analyze tailpipe emissions. Some researchers focused on point representation of cars and concentrate on their actual physical characteristics to build generalized linear models or stochastic models parameterized by air resistance, rolling resistance, and other physical aspects (Fontaras & Panagiota, 2011). Alternatively, others built a simplified linear model using variables such as the car mass, engine output, and fuel type. They showed that smaller passenger cars using diesel had less overall CO2 emissions than their diesel or larger European counterparts (G. Mellios, 2011).

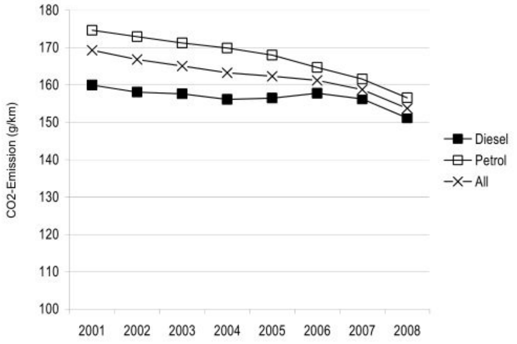


Figure 2 European ICE emissions (G. Mellios, 2011)

Researchers have also shifted from a single-engine point car model to more aggregate systems that consider traffic conditions or city characteristics as a whole (Kevin R. Gurney, 2012)—one such study reinforced that ICE engine CO_2 emissions account for 90% of macro emissions within most cities (Toshiko, 2003). (J.A. Paravantis *, 2006) incorporated changes in the overall automotive fleet in Greece as a baseline for CO_2 reduction. (Kii, 2020) took the societal aspect of emissions even further by incorporating changing population dynamics and road infrastructure into CO_2 emission modeling. They concluded that technology and a slower-growing population would drive emission reductions and thus less transportation and emissions.

Methodology

We initially set out to create a meaningful multiple linear regression model that could estimate the amount of carbon emissions (g/km) for a wide array of ICE vehicles in Canada. To do this, we had first to identify features that account for the variance in these emissions. Linear regression models offer insight into the relationship between the input variables (x) and the single output variable (y). For our purposes, this was an effective initial technique to determine which factors were correlated with carbon emissions from ICE vehicles and estimate our single output of carbon emitted (g/km) given different vehicle features.

The data used to identify these features and help build our multiple linear regression model was initially released on the Government of Canada website and then posted on Kaggle (Podder, 2020). We quickly found that the dataset included a large proportion of car makes and models built by companies such as Ford, Chevrolet, BMW, and Mercedes Benz, but also included lesser-known companies that sell to niche markets, including Bugatti and Smart. Additionally, there was a wide variety of vehicle classes, with the most common being classified as small sport utility vehicles (SUVs), mid-size, and compact vehicles. Fuel type was quite ubiquitous across vehicle classifications, with most either requiring regular or premium gasoline. However, there was a small percentage that ran on ethanol or diesel. Most vehicle transmissions fell under the category of “automatic select” or “automatic”; however, there was also a relatively large proportion of manual transmissions with 5 or 6 gears. Most cars contained 4-cylinder engines, but there was a range from 2 to 16 cylinders.

When evaluating the kurtosis of our continuous variables in the dataset, we found that fuel consumption was slightly right-skewed. Therefore, we subjected these features to Box-Cox transformations to normalize values, where our lambda value ranged from -0.4 to 0.2. Based on these outputs, we created more normalized distributions for these features, which helped provide flexibility for our linear regression models.

We combined these newly transformed variables with our original nominal variables after we subjected our data to feature transformations. With one compiled dataset, we split the data into a training and testing set – 80 percent was denoted to the training dataset, and 20 percent was designated to a holdout testing set.

Initially, we noticed from a preliminary linear regression model that fuel consumption (in miles per gallon) and engine size (in liters) were quite predictive of carbon dioxide emissions. Therefore, we added additional vehicle class features, cylinder engine composition, transmission type, and fuel type to the model, subjected it to stepwise selection, and yielded an R^2 value of about 0.98.

Experimentation and Results

We compiled summary statistics on our data set to understand the data before modeling and obtain distribution profiles of each variable's variables. Figure 4 represents histograms that describe the shape of the data. We noted that the features were almost normally distributed but were skewed right for all of the variables. We performed Boxcox transformations to normalize these distributions and better satisfy linearity assumptions associated with linear regression models.

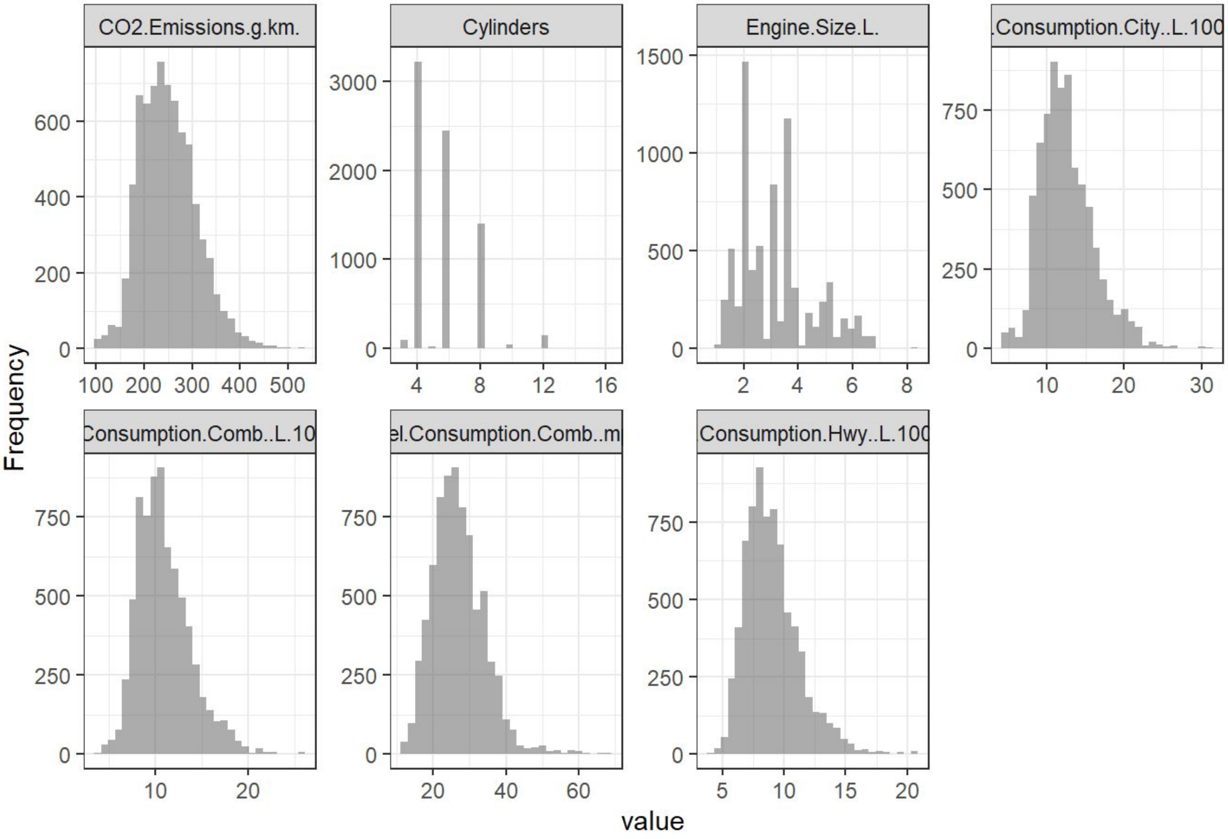


Figure 4 - Distribution of Numeric Variables

Additionally, four variables were considered non-numeric. These parameters are *Make*, *Vehicle Class*, *Transmission*, and *Fuel Type*. Figure 5 presents a bar plot of these variables and their prevalence in the dataset. We observed that Ford, Chevrolet, and BMWs were amongst the top 3 car manufacturers prevalent, and small to mid-size SUVs were the most common vehicle class. 6-speed and 8-speed automatic transmissions were heavily present and fuel type "X" and "Z" are the most common. We apply "dummifying" techniques to convert the categorical data into numeric features with 0's and 1's so these could be considered in our linear regression.

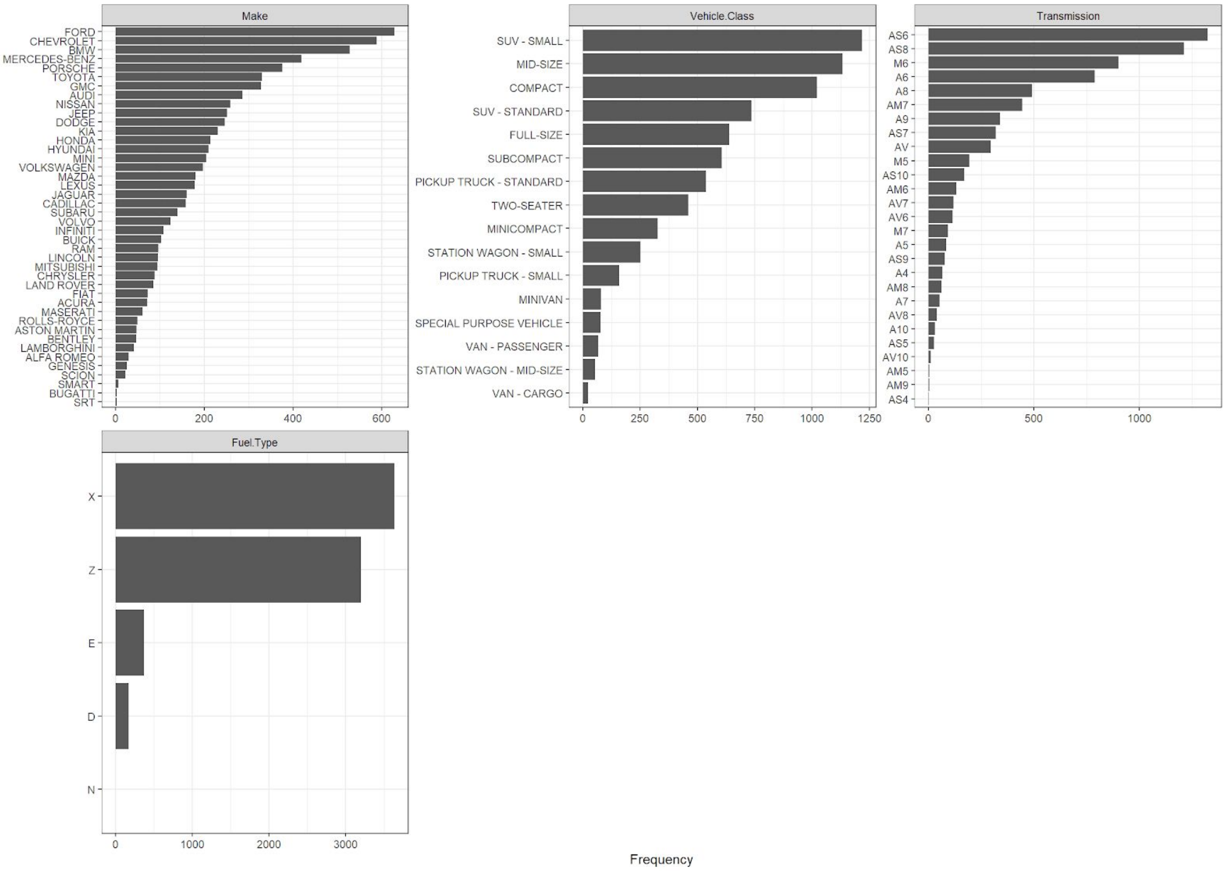


Figure 5 - Quantitative description of Qualitative Variables

Next, we obtain a preliminary assessment of our feature relationships to our target variable, CO_2 emissions. Figure 6 shows the individual relationships between CO_2 emissions and each of the numeric variables within the dataset. We observed that all numeric variables show some monotonic relationship with CO_2. Some of the relationships appear linear such as *Cylinders, Engine Size, Fuel Consumption City,* and *Fuel Consumption Hwy.* A relationship with *Fuel Consumption Combined* is also present in the form of what appears to be a polynomial-like function. Relationships like this require transformations to satisfy the assumptions of linearity in the regression problem. Additionally, Figure 7 shows a correlation plot that identifies relationships between all numeric variables. This is necessary to identify potential multicollinearity within the dataset. The presence of multicollinearity can increase the variance, also known as *variance inflation*. A variance inflation factor can be calculated for the dataset to potentially eliminate variables that have high VIFs.

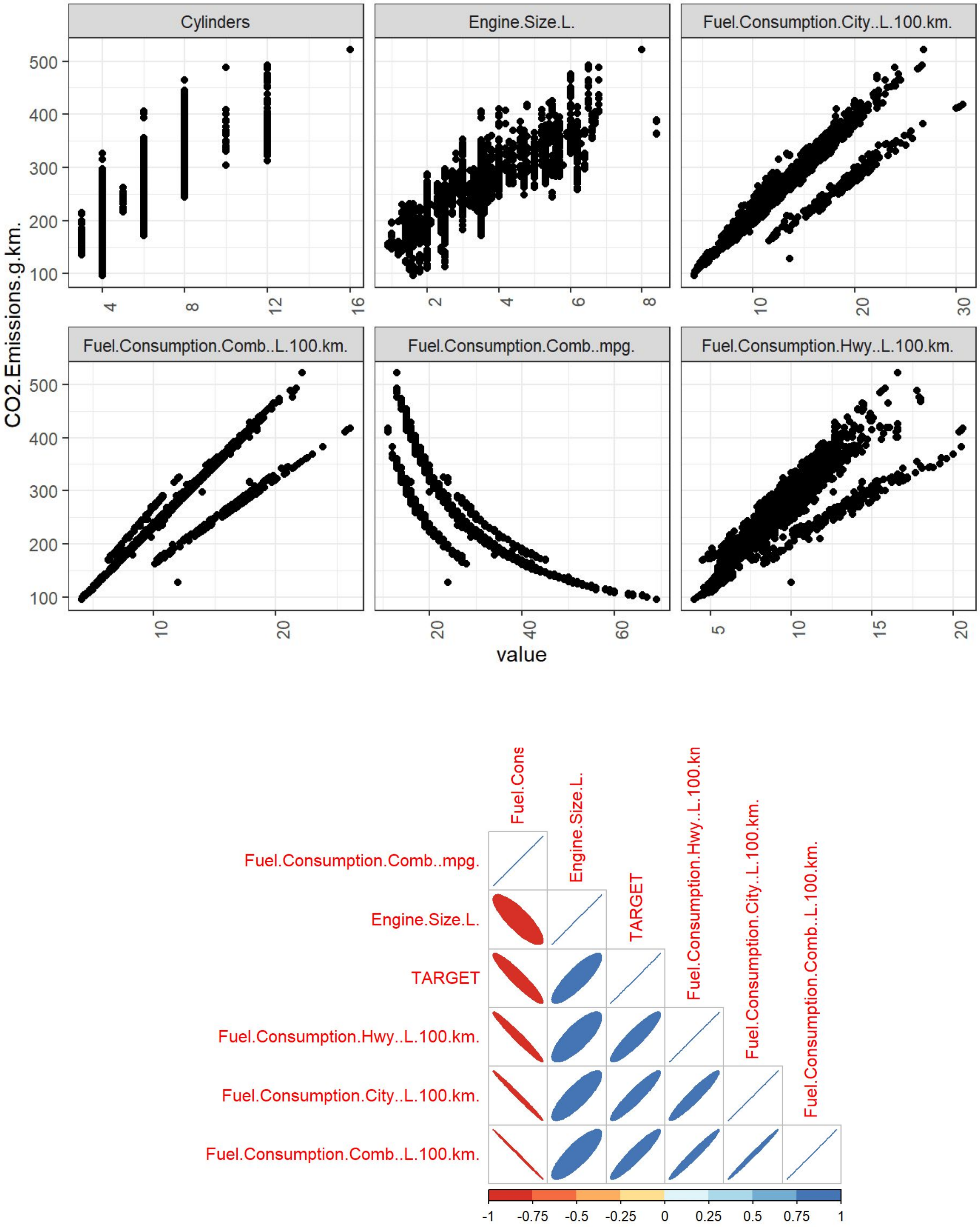
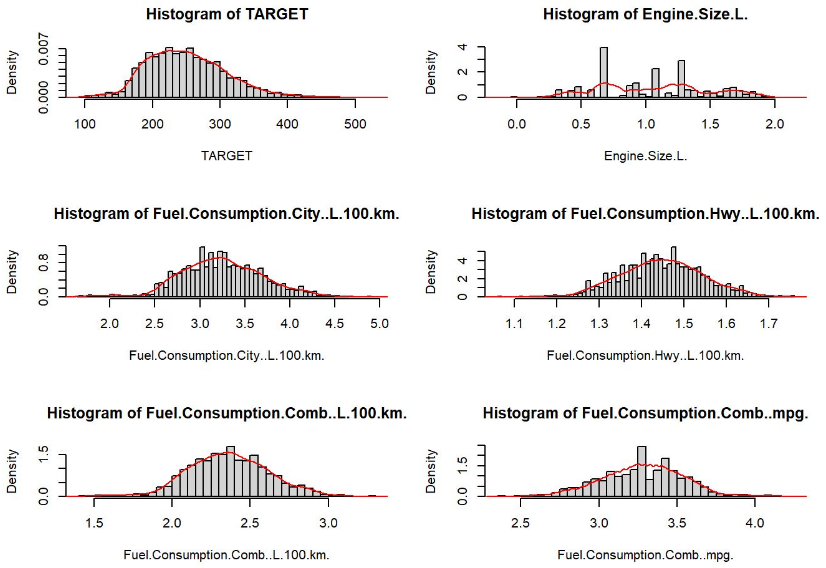
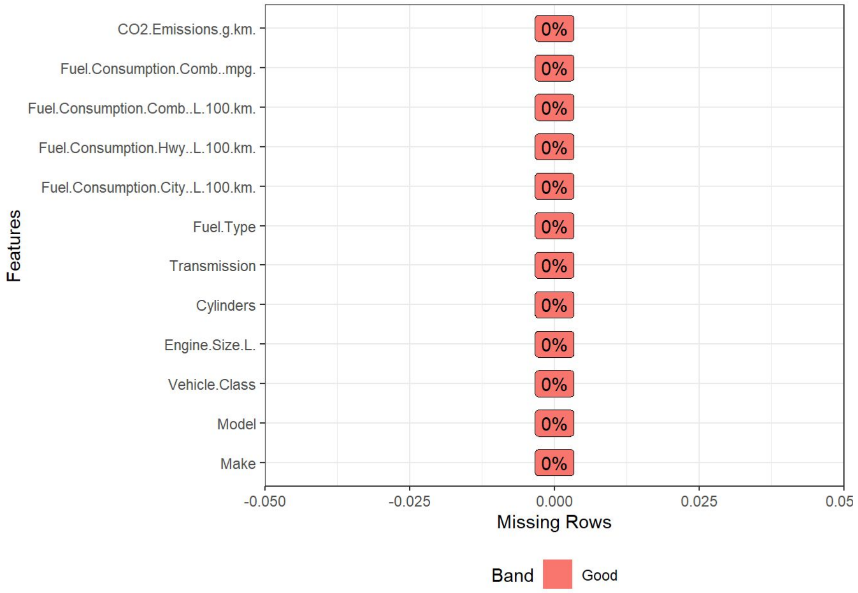


Figure 7 - Numeric Variables vs. CO$_2$ *Emissions*

Box-cox transforms with optimal lambdas determined by R were used to transform all of the dataset variables. Figure 9 shows these transformed variables, and in comparison to the untransformed data, these appear to follow a more gaussian-like distribution. Finally, imputation on the dataset for missing information was not needed, as shown in Figure 8. This figure represents the amount of sparsity within the dataset, and fortunately, this Kaggle dataset did not have any missing information, and no imputation was necessary.





Discussions and Conclusions

Model	RSE	Adj. R2	F. Statistic
Model 1	19.32	0.891	1.207x10^4
Model 2	7.577	0.983	6830
Model 3	7.577	0.983	6967

Table 1 Model Results

The results were consistent between the training and the testing sets for both model 2 and model 3. Because of the simplicity of model 3 with StepAIC and the reduction of variables considered via stepwise selection, no compromise to the performance was made except for a slightly higher *F-statistic* implying slightly higher statistical significance although negligible. Train/test evaluation was not done for model 1 since models 2 and 3 performed significantly better on the training sets.

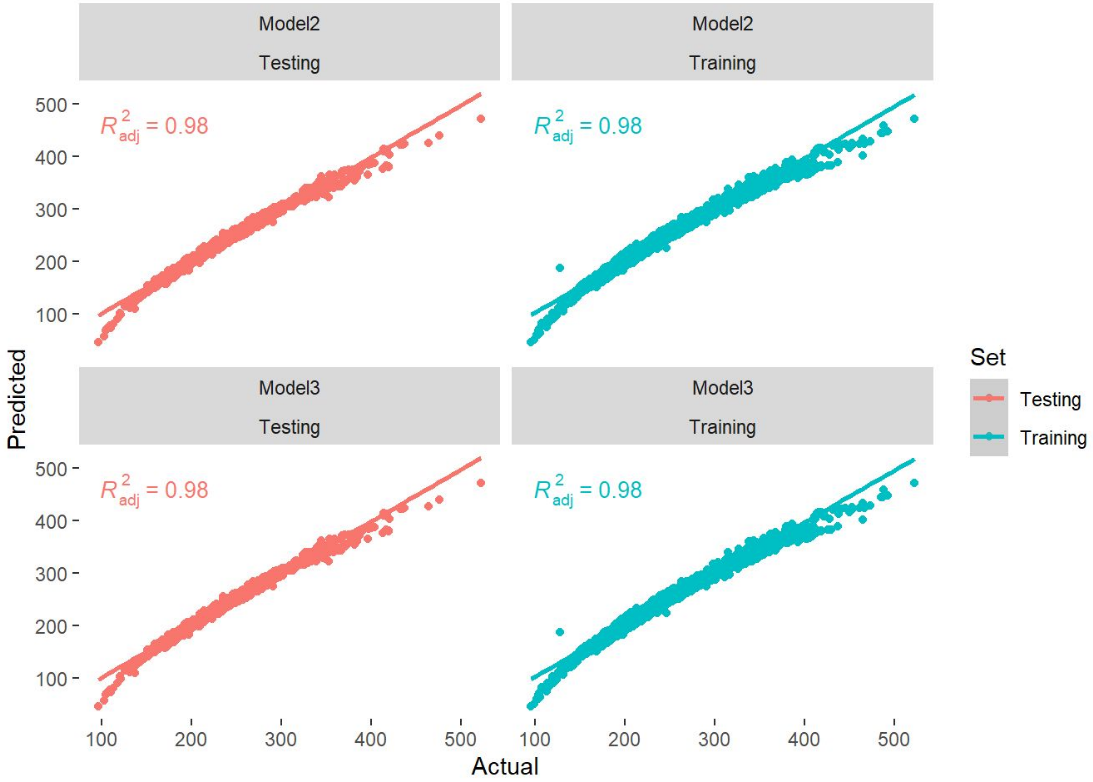
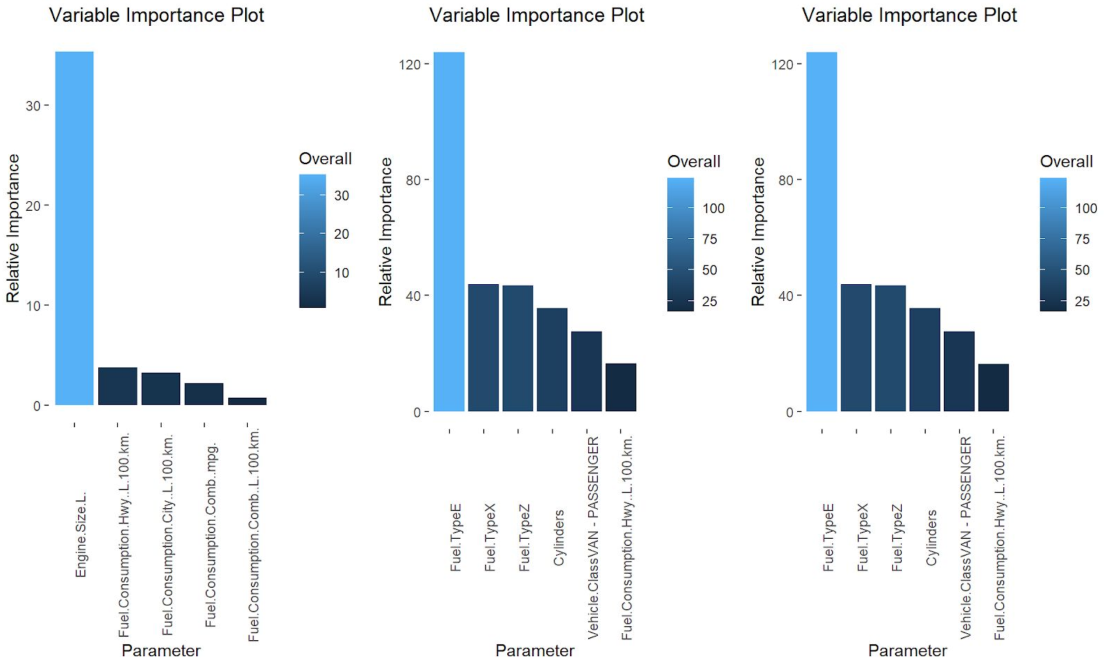


Figure 10 - Test/Train comparison for Model Performance

Finally, Variable importance was observed for all variables, and Figure 11 shows the top 5 predictors of each model. Model 1 differed significantly in the variables that were most important in predicting CO_2 emissions. The largest predictor was engine size; however, for models 2 and 3, they consistently determined that specific fuel types were the largest determining factor of CO_2 emissions.



References

1. (2019). *Clean Fuel Standard.* Environmental Change of Canada. Retrieved from https://www.canada.ca/content/dam/eccc/documents/pdf/climate-change/pricing-pollution/Clean-fuel-standard-proposed-regulatory-approach.pdf
2. Fontaras, G., & Panagiota, D. (2011). The evolution of European passenger car characteristics 2000–2010 and its effects on real-world CO 2 emissions and CO 2 reduction policy. *Energy Policy*.
3. G. Mellios, S. H. (2011). Parameterisation of fuel consumption and CO 2. *JRC Scientific and Technical Reports*.
4. *Government du Canada*. (n.d.). Retrieved from Energy and the economy: https://www.nrcan.gc.ca/science-data/data-analysis/energy-data-analysis/energy-facts/energy-and-economy/20062
5. Hope, C., & Alberth, S. (2008). *The Cost of Climate Change: What We'll Pay if Global Warming.* NRDC.
6. J.A. Paravantis *, D. G. (2006). Trends in energy consumption and carbon dioxide emissions. *Technology in Forecasting and Societal Change*.
7. Kevin R. Gurney, I. R. (2012). Quantification of Fossil Fuel CO 2 Emissions on the Building/Street. *Environmental Science and Technology*.
8. Kii, M. (2020). Reductions in CO 2 Emissions from Passenger Cars in Japan under Population and Technology. *Sustainability*.
9. Molico, M. (2019, November 19). *Researching the Economic Impacts of Climate Change*. Retrieved from Bank of Canada: https://www.bankofcanada.ca/2019/11/researching-economic-impacts-climate-change/
10. Nadia, P. (2020, Feb 20). *Climate Change Rises as a Public Priority. But It's More Partisan Than Ever.* Retrieved from New York Times: https://www.nytimes.com/interactive/2020/02/20/climate/climate-change-polls.html
11. Shah, M. (2019, October 9). *Climate change emerges as one of the top ballot-box issues among voters*. Retrieved from https://globalnews.ca/news/6006868/climate-change-federal-election-issue-poll/
12. Tabuci, H. (2018, April 2). *New York Times*. Retrieved from Calling Car Pollution Standards 'Too High,' E.P.A. Sets Up Fight With California: https://www.nytimes.com/2018/04/02/climate/trump-auto-emissions-rules.html
13. Toshiko, N. (2003). Energy Modeling on Cleaner Vehicles in Japan. *Journal of Cleaner Production*.